Differential Privacy 201 and the TopDown Algorithm

May 13, 2021
1 pm ET

Coordinator:
Welcome everyone and thank you for standing by. I would like to advise you that today's call is being recorded. If you have any objections you may disconnect at this time. Also all participants will be in listen-only mode for the duration of today's call. I would now like to turn the conference over to Michael Hawes from the US Census Bureau. Thank you. You may begin.

Michael Hawes:
Thank you operator and good afternoon everyone. Welcome to the third Webinar in our series on Understanding the 2020 Census Disclosure Avoidance System. Today's session will provide a deeper discussion on some of the concepts underlying our implementation of differential privacy in the 2020 Census Disclosure Avoidance Systems TopDown Algorithm.

I'm joined today by my colleague Michael Ratcliffe, Senior Advisor for Frames in the Census Bureau's Geography Division who will be speaking in a little while about the TopDown Algorithm's geographic hierarchy. Also on the line are my colleagues, Meghan Maury, Michele Hedrick, Philip Leclerc, Pavel Zhuravlev and Ryan Cummings who will be answering your questions during the presentation via WebEx's Q&A feature.

So as you come up with questions just type those into the Q and A. Please send them to all panelists and they will be responding to them in real-time during the presentation today.

Before I start I want to thank my Census Bureau colleagues and our external partners who have contributed to the information in this presentation. I'd also like to thank the many external

stakeholder groups who have provided invaluable feedback that has helped us improve the Disclosure Avoidance System over the past few years. I'd also like to state that any opinions and viewpoints expressed today are entirely my own and do not represent the opinions or viewpoints of the US Census Bureau.

The 2020 Disclosure Avoidance Systems TopDown Algorithm or TDA for short will be the method used to implement differentially private noise infusion for the first set of 2020 Census data products which include the public law 94-171 redistricting data, the demographic profiles and the demographic and housing characteristics files. The TDA will also be used for any special tabulations of the 2020 Census.

There are several requirements that TDA has been designed to meet. First the system must be able to ingest the Census Edited File microdata and geographic reference file and must output the microdata detail file that will feed into Decennial Tabulation Systems.

Next the algorithm must be able to hold certain data elements exactly as enumerated. We call these invariants and the list of invariants set by the Census Bureau's Data Stewardship Executive Policy Committee or DSEP for short includes total population at the state level, the number and type of occupied group quarters facilities at the block level and the number of housing units whether occupied or not at the block level.

In addition to these invariants TDA must be able to constrain values according to certain edit rules that were enforced on the Census Edited File. These include structural zeros like not allowing 3-year-old grandmothers as well as rules that require each occupied group quarters facility to have at least one occupant, among many others.

Perhaps most importantly the system must allow DSEP to determine the overall balance between privacy protections and the resulting data's fitness for use along with the ability to

prioritize accuracy across different tabulations at different levels of geography. And we'll talk more about this in a few moments.

The system must be able to ensure that the selection of privacy-loss budget directly controls the resulting accuracy of the data. Essentially that as you increase the privacy-loss budget to infinity the algorithm will eventually output the original CEF microdata.

And lastly transparency. All of the design code and parameters of the Disclosure Avoidance System must be able to be made public.

At a high level the TDA has five steps. It inputs the Census Edited File microdata and the geographic reference file geocoding. Then it converts those microdata to a functionally equivalent histogram of counts. You can think of this as a fully saturated contingency table of every variable crossed with every other variable with the value of those cells being the number of people at that level of geography who have those specific characteristics, along with all structural zeros which will be impossible values according to the edit rules for the CEF having been removed.

Then the algorithm asks a number of queries and injects noise into those results. We call this the noisy measurement stage. And I'll talk more about this step in a moment.

Armed with these noisy measurements the system must then perform a set of optimization problems. These are designed to ensure consistency across tables and geographies and to ensure that the final histogram is populated with non-negative integer counts.

Finally, the algorithm transforms the resulting histogram back into privacy protected microdata that can be output into the Decennial Tabulation Systems.

I mentioned that the first stage of the TopDown Algorithm is the conversion of the confidential Census Edited File, or CEF microdata, into a histogram that is functionally equivalent and a complete representation of the microdata itself. To understand what this actually means here is a simple illustration. On the left we have nine microdata records including their geographic location and all of their reported characteristics.

To convert these into a histogram, you identify all possible permutations of the location and every characteristic and then count the number of records in the microdata file with that exact combination of location and characteristics.

This set of record frequencies is the histogram which is used to take the noisy measurements for the TDA processing. After all of the processing and post processing steps, which we will discuss in a moment, the end product of the TopDown Algorithm is a privacy protected version of this histogram which can then be fully converted back into microdata by writing individual microdata records for each of these record counts with their corresponding location and unique combination of characteristics.

These privacy protected microdata then feed into the Decennial Tabulation System to produce the official Census data products.

So let's dive a little deeper into some of the steps in the TDA process. The noisy measurement step is what protects privacy in the algorithm. Before TDA goes into production our Data Stewardship Executive Policy committee will be setting the privacy-loss budget for the redistricting data product and will determine the allocation of that privacy-loss budget across the different queries that support the PL file and across the different levels of geography at which those queries are performed.

With those allocations of privacy-loss budget in hand the algorithm adds noise to each query that is performed against the confidential data. The noise that is added is taken from a

probability distribution with a mean of zero and with a variance that is determined by the share of the privacy-loss budget allocated to that particular query at that level of geography.

These noisy measurements are all independent of each other so they may not be internally consistent and they can include negative values. That is, it might say that a block that actually has zero Native Hawaiian or Pacific Islander residents now has negative one Native Hawaiian or Pacific Islander residents.

In order to meet the microdata output requirement of the TDA these noisy query answers need to go through postprocessing to make them internally consistent and non-negative. And we'll be discussing this post processing step in a moment. But first let’s look a little more closely at noise probability distributions.

Using our histogram illustration from a moment ago, imagine you wanted to ask three queries of the confidential data for this particular block. What is the total population? How many males are there, how many females? And how many members of each reported racial group?

Well to take the noisy measurements the TDA sums up the corresponding number of records to satisfy that particular query. Then it adds or subtracts a small amount of noise to the total. In this example you’ll see that three of the queries received zero noise. That’s because the probability distribution from which the amount of noise is randomly selected is centered on zero. So there is a highly likelihood of the algorithm adding no noise for any given query.

Plus or minus one occurs with the next greatest frequency. Plus or minus two with slightly less likelihood and so on. One important thing to notice in this example is that the different sets of queries, total versus male plus female versus the sum of the individual race categories give slightly different values for the block's population -- nine versus eight versus ten. These inconsistencies in the noisy measurements require postprocessing to make them consistent and we’ll talk more about how this postprocessing is done shortly.

Delving a little deeper into the mechanics of the probability distributions that we use to select the amount of noise added to the noisy measurements, in the fall of 2020 the Census Bureau made a change in TopDown Algorithm from traditional differential privacy, injecting noise to each statistic from the geometric distribution, which is the discrete equivalent of the Laplace probability distribution shown here in green, to concentrated differential practice by injecting noise from a discrete Gaussian distribution shown here in purple.

We implemented this change based on feedback we received from our data users who were concerned about the occurrence of significant outliers in the earlier demonstration data products that we had released. The notable distinction between geometric noise and discrete Gaussian noise is how those mechanisms handle the tails of the distributions.

Gaussian distributions have much flatter tails than their geometric counterparts meaning that any particular statistic would have less likelihood of receiving an unusually large amount of noise with a Gaussian mechanism that it would with the geometric mechanism for the same overall level of privacy protection.

So among other things this change helped us to reduce the occurrence of outliers in the resulting data for any comparable level of privacy. I should note for the sake of illustration, this image shows continuous loss in Gaussian distributions. But the TopDown Algorithm actually uses discrete versions of these distributions so only integer values of noise can be selected.

Our switch to concentrated differential privacy also improves the efficiency of the privacy-loss budget because zCDP composes better than traditional mechanisms of differential privacy. Essentially the privacy-loss budget goes further under zCDP.

Delving a little deeper into this change it's helpful to examine some of the differential privacy parameters and how they relate to privacy protection. Differential privacy is at its core a

framework for providing a mathematical guarantee of the maximum amount of confidential information leakage, privacy-loss or privacy risk, associated with publishing any statistic.

The switch from geometric noise to discrete Gaussian noise significantly reduces the likelihood of outliers yielding substantially greater accuracy for comparable privacy risk. It does so in part by modifying the mechanics of this mathematical guarantee.

We know that the publication of any statistic calculated from a confidential data source will inevitably reveal a small amount of confidential information in the process - privacy-loss. In traditional implementations of differential privacy, the privacy-loss parameter is represented by the Greek letter epsilon. The value of epsilon establishes the absolute upper bound on the amount of privacy-loss that can occur in shares of epsilon are allocated to each query and then sum, to the global value of epsilon for the data product.

In zCDP privacy-loss accounting is modified from the mathematical framework of pure DP and is quantified instead by the paired parameters epsilon and delta. By altering how the mechanism deals with unlikely events in the tails of the noise distribution, concentrated differential privacy using a Gaussian distribution has a probabilistic term delta to interpreting the mathematical guarantee represented by epsilon.

In this framework delta can be interpreted as the minuscule likelihood, such as one chance in 10 billion that the amount of privacy-loss might possibly exceed the upper bound established by epsilon.

In concentrated differential privacy, epsilon and delta always interact as a pair, and the same relative accuracy can be interpreted as many different paired values of these terms. For example, to meet a specific accuracy target the same noise distribution used to protect the statistics can be represented by many different epsilon delta pairs, a smaller epsilon with a higher delta or a larger epsilon with a smaller delta.

These different pairs of values hypothetically for example, an epsilon of four with a delta of ten to the negative six verses and epsilon of eight with a delta of 10 to the negative 10, represents exactly the same noise distribution but help you interpret your confidence in the upper bound of privacy-loss that's reflected by that value of epsilon.

With this example you would know that the probability of the privacy-loss possibly exceeding an epsilon of four is one in 1 million and the chance of it exceeding epsilon of eight would be one in 10 billion. In this regard selection of delta is a policy decision but only in so far as it determines how you're going to interpret an account for the mathematical privacy guarantee itself. It does not directly impact the resulting accuracy of the data. Currently the Census Bureau's privacy accounting uses a value of delta of 10 to the minus 10, so our published values of epsilon should be interpreted accordingly.

The other parameter introduced by (zCDP) is represented by the Greek letter Rho. In zCDP Rho is related to epsilon but is calculated differently. While in traditional DP, privacy-loss budget is allocated via shares of epsilon, in zCDP privacy-loss budget is allocated by shares of a parameter Rho.

These shares can then be added up and the global Rho can be converted into its corresponding value of epsilon for your chosen level of delta. And you'll see this in practice a little later in the presentation.

So now that we've discussed how the noisy measurements work, how do we deal with negative noisy values or values that don't sum consistently? Well this is done through post processing of those noisy measurements. And it's at the center of how the TopDown Algorithm operates.

So how does the algorithm perform this postprocessing? Well as the name suggests, we use a top-down approach. The algorithm starts by postprocessing the national level histogram then

moves its way down each level of the geographic hierarchy processing those histograms in turn until it gets all the way down to the individual Census block level.

At each geographic level, the algorithm takes the noisy query answers for that level of geography, the invariance which I discussed before, and the structural and rule-based constraints which I mentioned previously and then determines the internally consistent non-negative integer histogram that best reflects those noisy answers from the noisy measurements.

As the algorithm moves down the geographic hierarchy, the histogram values determined for the geographic level above it get added to the set of constraints within which the algorithm must optimize.

The design of the TopDown Algorithm has a number of important advantages over other possible implementations of differentially private noise infusion for these data. The recursive process down the geographic hierarchy ensures that the disclosure limitation error does not increase when you aggregate Census blocks to higher level geographies.

This would not be the case with the bottom up approach which would result in higher geographic levels having significantly greater amounts of noise. And this is a key feature of official statistics that we wanted to ensure TDA observed, that the accuracy of your statistics improve as the measured population size increases.

Lastly TDA has a helpful efficiency built-in insofar as the algorithm allows lower levels to inherit accuracy or borrow strength from the measurements taken at higher levels. This helps improve count accuracy at lower geographic levels without expending additional privacy-loss budget.

So I showed a moment ago how the TopDown Algorithm performs the postprocessing steps descending along a geographic hierarchy or a geographic spine as we often call it. It's important to note that TDA only takes noisy measurements for geographic units on the hierarchy, so accuracy for on-spine geographies will normally be higher than the corresponding off-spine geographies of comparable size.

But many legal and political geographies of interest are off-spine. Therefore their accuracy is impacted by the accuracy of the minimum number of on-spine geographies that could be used to construct them, that is by the number of on-spine geographic units that would need to be added or subtracted to construct the geography of interest.

To address this challenge the Disclosure Avoidance System Team made changes to the geographic hierarchy to improve the accuracy of off-spine geographical entities. This was done primarily through the creation of what we call optimized block groups whereby the algorithm reconfigures block group boundaries to bring these off-spine entities closer to the spine essentially minimizing the off-spine distance and also by isolating group quarters facilities from the surrounding areas.

As can be seen in the April 2021 demonstration data this optimization of the geographic spine significantly improves accuracy for off-spine entities and was central to our ability to tune the algorithm for the redistricting and Voting Rights Act use cases. I want to stress however that the optimization of the geographic hierarchy only impacts how TDA operates. It will not affect tabulation geographies in the published Census data products.

So, to explain how the optimization of the TDAs geographic hierarchy was done, I'll turn things over to my colleague Mike Ratcliffe, Senior Advisor for Frames in our Geography Division.

Michael Ratcliffe:

Thank you Michael. Can you could go to the next slide please? We'll talk a little bit about how we rethought the geographic hierarchy.

So in Michael's last slide we saw the central hierarchy in Census Bureau's geographic overall standard geographic hierarchy and the central axis if you want to call at that. Each of the types - and we'll see this again in the next slide that, you know, slide that follows. Each of the types of geographies in that hierarchy or that spine nests neatly within the higher level entity type. So it's a very nice neat aggregation of geographic units. So blocks add up the block groups, block groups to tracts, tracts to counties and so on all the way up to the US.

But there are of course as Michael mentioned, there are other geographic entity types that are important to data users. And these other off-spine entities intersect with the on-spine entities in myriad and sometimes complex ways. So the challenge to us -- and we've heard this in meetings with stakeholders, external stakeholders -- the challenge was to provide for the direct measurement of population and characteristics for American Indian and Alaska Natives, Native Hawaiian areas, and substate legal geographic areas when applying the differential privacy methods. There are of course many other geographic areas that we publish data for but these were the primary - these were the key types of geographies that we heard from our stakeholders that were critical and important.

So the consideration - so that was our challenge. The consideration was that - and Michael alluded to this, the larger the number of geographic areas on the geographic hierarchy or the spine and the more intersections between geographic areas that are formed when one type of area overlaps with another, the more thinly the privacy-loss budget is distributed impacting the accuracy of data for all geographic areas.

So our solution has been to bring the legal American Indian Alaska Native, Native Hawaiian areas on as well as places, incorporated places in Census designated places in 38 states and towns into cities and towns and townships in 12 states closer to the spine for DAS processing.

Now those of you who know our Census geography and our types of geography know that Census designated places are not legal entities. They're unincorporated communities, yet they are quite important in some states as place level entities. So when we say legal we'll use that with a little bit of poetic license and include the unincorporated places in 38 states. And in the American Indian Alaska Native areas we are also including the Oklahoma tribal statistical areas, which are the former reservations that existed in Oklahoma as well as the Alaska Native village statistical areas, which are the statistical representation of the legal existing Alaska native villages.

Next slide please. So to kind of recap our standard hierarchy we saw this again in Michael's last slide. But this is the usual view that that you'll see in our products and on our Web site. And you can see the central hierarchy there in the middle the central spine of blocks, block groups, tracts, counties or if you go top down, the nations, regions, divisions, states and so on.

I've highlighted in the box the geographic entities that are of importance in the DAS processing -- so states, counties, tracts and so on. And again as you can see there are many other geographic areas off the spine, off to the side that intersect in a variety of ways.

On the right is a conceptual view of the two pathways we're taking in optimizing the off-spine geographies. So we start at the US and then within each state we divide that state into the portion of the state that is within American Indian Alaska Native are Native Hawaiian areas. And in 2010 there are 36 state areas that fall into that category.

For 2020 there will be 38. There's a new off reservation trust land in Tennessee and a new off reservation trust land in Indiana. So those are the two additional states that you'll see with American Indian areas in 2020.

And then on the on non-American Indian areas pathway, we have 51 states or state-equivalent areas that or the portions of the states that are not within American Indian areas. And you can see that we follow that pathway down through counties and tracts through block groups, blocks, the optimized block groups that Michael was talking about.

Next slide please. So how does this work for the American Indians? What is the state portion, the state American Indian area that we're talking about? In this example we're looking at Kansas and there are three American Indian areas, four if you want to treat the off reservation trust land as a separate entity. But there are three American Indian areas in Kansas that are grouped together at the state level and then are used in the processing, in the postprocessing in the optimization.

So the Iowa, Kansas, Nebraska reservations and off reservation trust lands, Kickapoo Kansas Reservation and the Prairie Land, Potawatomi Nation Reservation.

Next slide please. Now before we turn to the non-American Indian Alaska Native, Native Hawaiian path I'd like to spend a little bit of time discussing the regional variations in incorporated places and Minor Civil Divisions. We took the variation in the primary units of local government into account in our optimization. And I talked about the 38 states where we're focused on place and incorporated census designated places and then the 12 states where we focus on Minor Civil Divisions -- cities, boroughs, towns and townships.

Those 12 states are what we refer to sometimes as the strong minor civil division states. And in this map they are the states in purple, in the darker shading of period, the nine states in the Northeast and then also the three upper Midwest states Michigan, Wisconsin and Minnesota.

In these states the Minor Civil Divisions are a category of county subdivisions in our census geography. In the Minor Civil Divisions, again these are cities, boroughs, towns and townships, the MCDs in the states have active functioning governments on par with the incorporated

places in other states. So these are the primary units of local government in those states. And so it was critical to take the MCDs in those states into account and distinguish them from the political geography that we see, the variety of political geography we see in legal geography in the other states. And those would be all of the other states and the other shades of purple, lilac and then the green.

All right, next slide please. So focusing on the geographic hierarchy on the more important substate geographic entities, so we focused on the geographic hierarchy on the more important substate geographic entities in recognition of the regional variations that exist.

I talked about the 12 strong Minor Civil Division states. In those we optimized block groups. The optimized block groups were configured to bring the Minor Civil Divisions closer to the spine. Again those are the primary units of local government. In those states they often are county subdivisions.

And in many instances the tracts and block groups will nest within those geographies. That's especially true in New England.

In the other 38 states the non-strong Minor Civil Division states, the District of Columbia and Puerto Rico, the optimized block groups were configured to bring places again closer to the spine.

And if you're familiar with our summary levels, that's summary level 160 state place. And again that includes not only the legal incorporated places but also the unincorporated Census designated places that we defined in cooperation with local officials. And I believe that's the end of my slides Michael so I'll turn it back to you.

Michael Hawes:

Great thanks Michael. In addition to our spine optimization there are a few other features of the TopDown Algorithm we're discussing. Back in 2019, as we were producing the first demonstration data product using 2010 Census data we noticed some unusual behavior in the TopDown Algorithm. Apparently the algorithm had difficulty effectively performing the post processing for large queries with a substantial number of zeros or small values.

The sparsity of the histogram resulted in some substantial biases in the resulting processing. Effectively we were shifting people from urban centers to rural areas and from large population groups to smaller ones.

To mitigate this effect, we modified the algorithm in a number of ways but most importantly - we modified it to be able to run in a series of passes rather than performing all of the optimizations for each geographic level at once. This allows the algorithm to optimize certain query sets and then use those results as constraints for the subsequent processing of the subsequent queries.

In the context of the redistricting data, this is relatively straightforward. At most geographic levels the TopDown Algorithm processes the total population counts first and then processes the remaining queries to be consistent with those population counts.

For larger data products that have even more sparsity, like the demographic and housing characteristic files, the number and order of these passes can help to further diminish the impact of sparsity on the algorithm and allows us to prioritize accuracy for certain tabulations.

The TopDown Algorithm is remarkably flexible and the algorithm can be finely tuned to meet various accuracy targets. This tuning is largely done through the allocation of privacy-loss budget by geographic level and by query. Here you'll see the allocation of privacy-loss budget by geographic level that's reflected in the April 2021 demonstration data release.

In addition to the global Rho, epsilon and delta parameters for the person's file on the left and the units file on the right, you can also see how the shares of privacy-loss budget have been allocated across the different geographic levels within the two files.

You can see that the tuning we performed to meet our accuracy targets for the redistricting use case allocated a substantial share of privacy-loss budget, or Rho, to the block and optimized block group levels.

And in addition to allocating privacy-loss budget by geographic level the TopDown Algorithm allocates shares of privacy-loss budget to each of the queries performed in the noisy measurements stage at each geographic level.

As you can see here, the detailed query which is the full cross of the household and group quarters, element by voting age by Hispanic by CenRace gives a sizable share of the privacy-loss budget at all geographic levels.

You'll also see higher allocations of privacy-loss budget to the total population query at the optimized block group level and county level and to the special query for total population at the state level that supports the separation of tribal areas in the optimized spine, which Michael Ratcliffe discussed a moment ago.

Remember the total query, which is the total population counts is held in variant at the national and state levels. So no privacy-loss budget is expended on this query at either level.

The Rho allocation assigned to total at the state level in this table is actually the amount of privacy-loss budget assigned to the state level queries for the total population of all American Indian and Alaska Native tribal areas within the state, and for the total population of the remainder of the state for those 36 in 2010, or 38 in 2020, states that include American Indian Alaska Native tribal areas.

And before I open the floor for my colleagues to answer some more of your questions, I'd like to encourage you to join us for the remaining Webinars in this series. Tomorrow, May 14, I will be highlighted the detailed summary metrics from our April 2021 demonstration data release. And next Friday, May 21, we will be providing an analysis of our empirical assessments of the recent demonstration data for the redistricting and voting rights act use cases.

If you would like to stay updated on our development of the 2020 Disclosure Avoidance System please subscribe to our newsletter. We send out updates every couple of weeks with the latest information on the development and implementation of the DAS. And if you'd like to learn more about our modernization of Disclosure Avoidance Methods for the 2020 Census, check out our Web site. We have a wide array of useful resources including frequently asked questions, fact sheets, videos, blogs and much more.

And with that I am going to introduce Meghan Maury who will help moderate some of your remaining questions and my colleagues and I will be happy to answer them for you.

Meghan Maury:
Thank you so much Michael. Hello everyone. I'm Meghan Maury and apologies for showing you my messy office. I couldn't get my virtual background to load.

Let's Michael - Michael and Michael this was extremely technical Webinar so I'm going to ask you a couple of questions that came in in the chat or that I think will help people sort of disentangle some of that more technical language. So first, baseline question, does the move from differential privacy to this zero concentrated differential privacy does it increase the risk that people's data can be re-identified?

Michael Hawes:

So that's a great question and I'll start but then I would love to pass it over to Philip or Pavel to chime in as well. I guess the question is does it increase risk compared to what?

The switch to zero concentrated differential privacy still allows us a full privacy accounting. It just changes the mechanism by which that accounting takes place.

So instead of just considering your value of epsilon, you have to consider that pair. And within that pair that delta element helps you understand like at what point that privacy protection might degrade. If you set delta infinitesimally small or set it to zero you would be back in the world of essentially of traditional differential privacy.

So what this does is it just essentially tells you what's your confidence in the privacy-loss guarantee reflected by any particular level of epsilon. Philip or Pavel, do you want to add to that?

Philip:
I actually thought you did an excellent job addressing it Michael. So what you're comparing to is the principal question. If you're comparing zCDP to pure differential privacy, zCDP is a qualitatively slightly weaker guarantee in terms of privacy. But it's tunable in the same way and as you push those parameters closer to the extreme values that Michael mentioned you get a guarantee that is sort of increasingly similar to the pure differential privacy one. And in any case if you're using zCDP or pure differential privacy in either of those two worlds you're getting a guarantee that just didn't exist with methods that predated these.

Meghan Maury:
Thanks for that. That's helpful. I'm going to jump to a couple of really simple questions one, a couple folks have asked, "How do I find these slides?" And so we will make sure that in the chat you can see the link to where all the slides for this Webinar as well as all the other Webinars in this series are available. We're also making the recordings available.

It does take us a couple days to get that information up on the Web but please circle back if you'd like to review this or take a closer look at the slides.

Another relatively easy question someone asked, "Is the most recent PPMF, the most recent demonstration data, the first demonstration data that uses this zero concentrated differential privacy or have we seen that in prior demonstration data?"

Michael Hawes:
So no, this is not the first that uses it. We actually did make the switch to the discrete Gaussian distribution in the fall. I don't off the top of my head recall whether it was for the September or November release. But yes, no we did make that shift before but it was still at the lower privacy-loss budget that we've been holding consistent across the first four demonstration data releases. So this is the first use of the new mechanism with the higher privacy-loss budget.

Meghan Maury:
Got it. Thank you for that.

I think this is a really interesting question. So one of the questions says with the TopDown Algorithm are geographies that are closer to the top of the spine assumed to be more accurate regardless of the size of the population? So, I think the question is if you have a block group that has, you know, 3,000 people in it or let's say 10,000 people in it and a tract which is higher up the spine that has slightly less people in it, will that tract sort of be - assumed to be more accurate than the lower level geography with more people in it?

Michael Hawes:
So that's not exactly a straightforward question to answer as it might seem. The better way to think of it is the absolute accuracy in terms of the count accuracy for any geography is largely going to be determined by two things, the distance from the spine, so on the spine entities will

at any level geography will likely have higher count accuracy than the ones that are farther from the spine.

And the other is the allocation of privacy-loss budget that's dedicated to that particular query at that particular geographic level. So the count accuracy is going to vary depending on those two things. But you can also think about accuracy in terms of relative accuracy.

And because we're tuning on count accuracy through the system, the relative accuracy of any given calculation at any given level of geography will of course be more accurate in relative terms like when the underlying population increases. The same absolute count difference becomes a much smaller percentage difference when you're dealing with larger populations.

Meghan Maury:
Got it. That's very helpful. Thank you. There's a couple questions here, and I know that you don't have state specific information at hand, but a few folks have been asking, you know, they're doing their analyses of the most recent demonstration data, still seeing a shift from urban to rural areas. Some people talk about specific states in here but maybe you can talk generally about whether the algorithm was able to kind of completely address the shift from urban to rural or if people should still expect some shift there?

Michael Hawes:
So this is essentially a consequence of one of the requirements that was baked in from the TopDown Algorithm early on. And that was the requirement that we output internally consistent micro-data into decennial tabulation. That requirement means that there will inherently be at least a minimum level of bias, if you will, because of the fact that you can't report negative numbers. That when you have a location that has a noisy measurement - a query that would result in a zero or a one and you add noise to it, that noise could be negative and that could pull the noisy answer into the negative.

And so because we can't report out negative values we have to bring those back up to zero, there's an upward push on low areas and a corresponding downward push on large areas and upward for small groups and downward for large groups. Now we've done a number of improvements to the algorithm over the last year and a half that have mitigated that to a very large degree.

So if you look at the April demonstration data you'll see there is significantly less of that type of distortion than there was in the prior demonstration data particularly in the October 2019 files. But it's probably not completely gone.

But if you are, as you're analyzing the files if you do notice distortions like that that are of particular note or that are particularly worrisome, bring those to our attention because there are certainly additional things that can be done to mitigate those impacts and like let us know when and where you see them and we'll see what we can do with those.

Meghan Maury:
Yes thank you. And that's a great point. All of the questions in here that relate to analyses that you're doing now, we really want to see all of those so we can see what we can do to address your concerns. So please…

Michael Ratcliffe:
And if I can just…

Meghan Maury:
Oh go ahead.

Michael Ratcliffe:
Since urban and rural were mentioned, if I can just put a plug-in. We have published proposed criteria for defining urban areas for the 2020 Census based on 2020 data. The public comment

period is open for another seven days through May 20. So if you're interested you can get in touch with me off-line or we have information on our Urban-Rural page on census.gov and we would love to hear thoughts and comments on our proposals.

Meghan Maury:
Yes thanks for that. I'm becoming more and more of a geography fan every day so it's definitely encourage folks to dig it on that.

There's a couple questions here about uncertainty metrics. And I think in particular people are saying are you going to cover uncertainty metrics in a future Webinar or is there a place that they can go to learn more about what kind of confidence intervals will be available to data users so they can best interpret the data? Can anyone speak a little bit to what the plan is for that?

Michael Hawes:
So I can speak to the issue but not necessarily to the plan yet. So one of the real advantages of the differentially private approaches to disclosure avoidance over traditional approaches like the swapping mechanism that we used in the past is that we can be fully transparent about what the impact of the mechanism is on the resulting data.

In prior decades, disclosure avoidance absolutely had an impact on the accuracy and fitness for use of the data. But quantifying that impact had to be kept secret because you had to protect your swap rates and protect the assessment of the impact on accuracy. You had to keep that confidential to prevent reverse engineering of the confidential data.

With formally private methods like the TopDown Algorithm, we can be completely transparent about pretty much every aspect of the implementation including all of our code, all of our parameters and the exact noise distributions from which noise is selected and the privacy-loss

budget allocations that determine those. From all that information you can calculate margins of uncertainty for any statistic that's published. You can build that into your analyses.

Now the Census Bureau has committed to providing our data users with guidance on interpreting the fitness for use of the 2020 Census data products. What form of that guidance on interpreting fitness for use will take is still being worked out so I can't say whether it's going to be through the publication of margins of uncertainty or some other mechanism. Those details are still to be determined.

But there are a number of ways that that kind of fitness for use guidance could be developed either by us or by those outside the Census Bureau. And as we work out the details of what that guidance will actually look like we'll be sharing that publicly.

Meghan Maury:
Yes thank you so much. I know that lots of folks are hungry for more information on that so I appreciate your kind of pointing the pathway to how we'll be talking that through.

There's a couple of questions here on how you quantify distance from the spine. If there's a way to quantify distance from the spine in a way that's kind of predictive of what kind of error will be introduced by that distance from the spine. I know that's a technical and complicated question but I wonder if you could give any guidance on that?

Michael Hawes:
I will defer to Ryan on that because he knows the optimization process better than anybody so…

Ryan Cummings:

So kind of the most straightforward notion of distance from the spine is the minimum number of on-spine Geo units that you need to add or subtract from one another in order to derive the geographic extent of the off-spine entity. So we call that the off-spine entity distance.

And you can just think of it as just a distance metric, how far is the off-spine entity of away from the spine. And generally it's error increase as that metric for entities which would have a higher off-spine entity distance. So generally having to add or subtract a lot of these on-spine entities in order to derive the off-spine entity will increase the error.

Meghan Maury:
And is there a place where people can find more information about how far from the spine a geography is. Is there anywhere we have more information about that particular component?

Ryan Cummings:
I don't believe we've actually released anything yet on that. There is a document that's in the process of being reviewed right now that I believe will be available in the future though.

Meghan Maury:
Got it. I will say by the way there are so many questions in this Q&A, we are definitely not getting to them all. We have about eight more minutes before we have to wrap up.

So I just want to put in a plug, if you have questions that you don't get answers to on the Webinar today, first we really encourage you to come back to our other Webinars. Second you can always submit questions to us via our email inbox. Michael do have a slide with that information on…

Michael Hawes:
Yes give me one second please.

Meghan Maury:
…how to give that feedback?

Great. Michael will bring that all up for you all but let's keep going on the questions for as long as we can. Someone just asked for a repetition. Is more…

Michael Hawes:
Actually I don't have that here unfortunately. The email is 2020D-A-S short for Disclosure Avoidance System, so 2020das@census.gov.

Meghan Maury:
Great thank you Michael. And someone just asked for a quick repetition, and I just want to clarify this because I think it's important. Is more of the privacy-loss budget introduced into the smallest geographies? I wonder Michael if you could just run back to that slide that shows a little bit more of how the privacy-loss budget is allocated.

And one thing I want to correct in the question, it says the higher the privacy-loss budget the less accuracy is introduced to the smallest geographies? But I think if I'm right that's actually backwards from the reality. The more privacy-loss budget you have allocated to a place, the more accurate that geography is. Is that right. Am I saying that right?

Michael Hawes:
That's right. The more privacy-loss budget you throw at a particular query, in relative terms, the more accurate it will be compared to the other queries. So you'll notice that the substantial amount of privacy-loss budget that we allocated to the optimized block group level and the block level, that was done primarily to improve the accuracy of those off-spine entities which we've discussed before because in many cases those off-spine entities need to be constructed by adding or subtracting block groups or blocks.

Meghan Maury:
Got it. And this sort of allocation really kind of is what you mean by tuning, right? We've got a couple of questions in here about, what is tuning versus a criteria? Can you…

Michael Hawes:
So...

Meghan Maury:
...clarify that a little bit?

Michael Hawes:
Yes the idea there is that there's lots of different ways to assess accuracy or fitness for use or accuracy for what purpose, accuracy of which queries at which levels. And there's going to be lots of trade-offs involved in any Disclosure Avoidance System.

We've been using traditional methods but it's made more explicit in formal privacy with the explicit privacy accounting the differential privacy offers. So with those trade-offs you need to have kind of standards against which to evaluate those trade-offs.

The accuracy targets that we set and that we've discussed in previous Webinars were developed based on stakeholder feedback. And we've developed accuracy targets that we thought reflected those use cases as presented to us.

And then we tuned the parameters with those accuracy targets in mind. So essentially we experimented with different allocations of privacy-loss budget across geographic levels and across the query sets to find the ones that best capture or best met those accuracy targets that we develop as well as other accuracy targets that we had developed internally for other priority uses of these data.

And so the tuning is the adjustment of the parameters and dials of the system and your evaluation of the results of changing those parameters, that evaluation is done against whatever accuracy targets that you've set.

Meghan Maury:
Got it. Thank you. And very many thanks to Michael Ratcliffe for rapid-fire answering geography questions in the chat. Michael I wonder if you could just pop back to one of the slides that shows the spine, because we're getting a lot of questions about sort of what's on-spine, off-spine, just getting that clarity again. And Michael can you just one more time just walk us through what are the what are those on spine geographies versus off-spine geographies?

Michael Ratcliffe:
Sure. Yes so when we talk about the on-spine geographies what we're talking about are these geographic areas that are in that central hierarchy -- nation, region, division, state, counties tracts, block groups and blocks. And specifically when we're talking about spine in the context of Disclosure Avoidance System processing it states the counties, the tracts, the block groups and blocks. These are the geographies that nest neatly within each other.

So what that means is you'll never have a county that crosses a state boundary. You'll never have a census tract that crosses a county boundary.

One thing to note there are population thresholds, minimum, optimum, maximum suggested population thresholds for census tracts. Because of that aspect of the criteria, you will - when you have a county, so the optimum population for a census tract is 4000 people, when you have a county and the minimum is 1200, when you have a county of less than 1200 people you're only going to have one census tract. We will not cross a county line to make that census tract larger to that optimum population. So there's a constraint there. There's a population criterion and then there's a spatial criterion that kind of operates in some ways as a constraint.

The off-spine geographies are all the other geographies that are to the right and left of that central hierarchy. And these geographies are going to - they block that up to all of those. So when we're defining our - actually when we're preparing our final geography for the census for tabulation we compile the boundaries, we update the boundaries of all of the higher level geographic areas -- block group tracts, places, counties you name it -- all the ones off to the sides.

And then we define the census blocks so the census blocks are the last geographic unit defined because they have to adhere to all those boundaries, all of the network, the latticework of boundaries and roads and other features that exist at higher levels.

But you can always be assured that a census - that census blocks will add up to every other geographic unit. All the other geographies are going to get split in some fashion. All the other on-spine geographies will be split in some fashion by the off-spine geographies. So a place can be ...

Meghan Maury:
Got it.

Michael Ratcliffe:
...if you follow the lines, place can be in two counties. Can never be in two states, but it can sometimes be in two counties. It can split tracts in multiple ways and so on and so on. To get to the question in the chat about enumeration, areas we actually don't use that term anymore, haven't for several - for quite a few decades.

We do talk about collection geography. And starting in 2000 we completely separated our concepts of collection geography and tabulation geography. So the collection geography exists only to carry out Census operations and then the tabulation and - then it's put aside. And the

tabulation geography, everything you see on the hierarchy chart is what's used to tabulate and then disseminate data.

Meghan Maury:
One clarifying question before we wrap up, I know we're right at time but you mentioned that those off-spine geographies are never going to cross state lines but there is an exception to that right? Tribal geographies might cross state lines?

Michael Ratcliffe:
Well I was referring to places. We sometimes get things, you know, we'll sometimes hear from people saying, "Well what about Kansas City because there are two Kansas City's. There are two Texarkana's and so on." Cities are creatures of the state.

But you're absolutely right. Tribal geographies will cross state lines. The Navajo Nation is in three states. Urban areas will cross state lines, metropolitan areas so, you know, if you follow the lines on the hierarchy chart you will see those relationships. It's not a perfect view. This is a conceptual view of trying to bring some order to a very complex and sometimes messy set of geographic units. But yes so tribal areas yes, will cross straight lines.

And so we've gotten questions about the Navajo Nation and Indian areas American Indian areas across state lines. So when we're applying the American Indian Alaskan Native, Native Hawaiian pathway here, the Navajo, the portion of the Navajo Nation in New Mexico will be grouped with all of the other American Indian areas in New Mexico. The portion in Arizona will be grouped with all the other tribal areas in Arizona and the portion in Utah and so forth.

Meghan Maury:
Got it.

Michael Ratcliffe:

And then it'll be brought back together into a single unit when we publish the data.

Meghan Maury:
Incredibly helpful and I really appreciate you clarifying that. I know it's a point that's been really tricky for people to understand. Thank you so much. I know we're at time. Michael Hawes do you want to take us out?

Michael Hawes:
Sure. So I just want to thank everybody again for joining us today and I'll go back to our schedule here. So please join us for the subsequent Webinars in the series. Again we have one tomorrow on the Detail Summary Metrics and one next Friday on our empirical assessment of the April 2021 demonstration data for the redistricting and Voting Rights Act use cases. Both should be great sessions.

If you'd like to see the recordings of the prior sessions in this series those are available on our Web site as we previously mentioned. And with that I will say thank you to everyone and hope you have a great rest of your day.

Coordinator: That will conclude today's conference and we thank you for participating. You may disconnect at this time. Speakers please stand by for your post conference.

END